

反向映射主成分分析方法在化工建模中的应用

秦 勇 沈静珠 胡山鹰 周建华

(清华大学化工系, 北京 100084)

周 进

(齐鲁石化胜利炼厂, 山东临淄 255434)

摘 要 基于多元统计中常用的主成分分析方法, 提出了反向映射主成分分析的数据处理方法及数学分析, 并通过建立重油加氢裂化反应器压降模型的实例将其与其他化工中常用的建模方法进行了比较。

关键词 主成分分析 反向映射 建模 神经网络 滴流床

引言

主成分分析(Principal Component Analysis, PCA)方法是多元统计中常用到的一类方法, 它具有去除数据相关性, 减少噪声影响, 降低数据维数等优点, 在过程控制、过程诊断、数据压缩、信号处理、模式识别等各个领域都获得广泛应用。80年代, S.Wold与H.Martens将其应用于化学计量学领域, 进入90年代, 开始在化学和化工中得到广泛的应用, 如其在化工动态过程中的应用, 另一方面又与其他的非线性方法如人工神经网络(Artificial Neural Networks, ANN)结合发展了非线性PCA和非线性偏最小二乘法(Partial Least Squares, PLS)^[1], 应用于稳态过程的建模, 并在动态过程的建模中获得了成功^[2]。

在建立数学模型时, 若将所有因素均引入, 无疑是困难的; 同时, 各个因素间又往往存在一定的相关性, 将大大增加分析和解决问题的复杂性。PCA是将过程中大量的相关信息和带有噪声的数据投影到含有所有相关信息的低维子空间上, 从而起到降维和去噪的作用。PCA降维实际相当于一个特征提取和信息压缩的过程。按照特征提取的观点, 应寻求一种映射(或变换), 将高维空间中的样本映射到低维空间中, 并使之与其他的变换映射相比是最优的。提取判断准则可采用基于距离的判据(如欧氏距离)、基于概率分布的判据(如概率距离度量或散度判据)、也可采用基于熵函数的判据(如Shannon熵, 平方熵等)^[3]。而PCA相当于一种基于最小均方误差的提取方法, 类似于Karhunen-Loeve变换。

1 反向映射主成分分析法

PCA方法可以描述为: 若矩阵 $X \in R^{n \times m}$, $\text{rank} X = r$, 则 X 可表示为

$$X = \sum_{i=1}^r M_i, M_i \in R^{n \times m}, \text{rank} M_i = 1 \quad (1)$$

矩阵 M_i 也可写成两个向量积的形式, $M_i = t_i p_i^T$, 故

$$X = \sum_{i=1}^r t_i p_i^T \quad (2)$$

式中 $t_i \in R^{n \times 1}$, 为坐标向量; $p_i \in R^{m \times 1}$, 为基向量。故

1997-10-28收到初稿, 1998-07-06收到修改稿。

联系人: 沈静珠。第一作者: 秦勇, 男, 23岁。

$$X = TP^T \quad (3)$$

式中 $P^T = (p_1^T, p_2^T, \dots, p_r^T)$, $T = (t_1, t_2, \dots, t_r)$, p_i 满足正交性

$$p_i p_j^T = \begin{cases} I & j = i \\ 0 & j \neq i \end{cases} \quad (4)$$

即通过 PCA 分解后, 原样本矩阵分解为一组标准正交基与该正交基所张成的空间中投影坐标的外积. 分解后的基和坐标用于建模、分类等. 求取 PCA 中各维主成分的方法很多, 常用的是 NIPALS(Nonlinear iterative partial least squares)方法^[4].

由求得的 PCA 各主成分即可用于建模, 其方法有主成分回归(Principal Component Regression, PCR)、PCA-ANN 等. 分析 PCA 方法可知, 若采用 PCA 分解后主成分中的前 a 维来代替原先的 r 维, 则可起到降维的作用; 同时由 PCA 方法分解原理可知, 各维主成分按 $X^T X$ 的特征值排列先后求出, 后几维往往代表误差信息. 用常用的 PCR、PCA-ANN 方法建模时通常可起到较好效果, 但对于 a 维主成分中含有误差时则无法准确建模. 而对各种工业过程, 建模所需数据大多来源于工业实际, 由于种种原因, 数据的准确性无法严格保证, 所以在作 PCA 分解时会产生主成分中即含有误差的情况, 这将严重影响所建模型的正确性.

为了克服主成分中含有误差对模型的影响, 本文提出反向映射主成分分析(Back Mapping PCA, BMPCA)的方法, 其基本原理是利用 PCA 中寻找到的主成分进行反向映射到 $\hat{X}$, 若取 r 个主成分中的前 a 个用于表征样本集中的主要信息, 则

$$\hat{X} = \sum_{i=1}^a t_i p_i^T \quad (5)$$

t_i 是在主成分 p_i 张成的子空间中的投影, 其中包含了 a 维主成分中的全部信息, 而这些信息是经过特征提取得到的; 再由其反向映射到原样本空间中, 使集中于某几维主成分的误差在原样本空间中重新分布, 减小了主成分中的误差. $\hat{X}$ 是 X 所存在空间中与 X 最接近的一组样本.

BMPCA 原理数学分析如下

$$X = \sum_{i=1}^r t_i p_i^T \quad (6)$$

由式(4)主成分的正交性, 则 BMPCA 的截断误差可表示为

$$\begin{aligned} \xi &= (X - \hat{X})^T (X - \hat{X}) \\ &= X^T X - X^T \hat{X} - \hat{X}^T X + \hat{X}^T \hat{X} \\ &= \sum_{i=1}^r p_i^T t_i p_i^T - \sum_{i=1}^a p_i^T t_i p_i^T - \sum_{i=1}^a p_i^T t_i p_i^T + \sum_{i=1}^a p_i^T t_i p_i^T \\ &= \sum_{i=1}^r p_i^T t_i p_i^T - \sum_{i=1}^a p_i^T t_i p_i^T \\ &= \sum_{i=a+1}^r p_i^T t_i p_i^T \end{aligned} \quad (7)$$

由投影关系容易知道

$$t_i = X p_i \quad (8)$$

令 $\Psi = t_i^T t_i$, 则

$$\xi = \sum_{i=a+1}^r p_i \Psi p_i^T \quad (9)$$

用拉格朗日乘子法, 可求出在满足正交条件下, ξ 取极值的坐标系统, 即用函数

$$g(p_i) = \sum_{i=a+1}^r p_i \Psi p_i^T - \sum_{i=a+1}^r \lambda_i [p_i p_i^T - I] \quad (10)$$

对 p_i , $i = a+1, \dots, r$ 求导数, 有

$$(\Psi - \lambda_i I) p_i^T = 0 \quad i = a+1, \dots, r \quad (11)$$

当取 a 个 p_i , $i = 1, 2, \dots, a$ 来逼近 X 时, 其截断误差

$$\xi = \sum_{i=a+1}^r \lambda_i I \quad (12)$$

故当取 a 个与矩阵 $X^T X$ 的 a 个最大特征值对应的特征向量来展开 X 时, 其截断误差和所有其他正交坐标系情况下用 a 个坐标展开 X 时所引起的截断误差相比最小. 因此用 $\hat{X}$ 代表原先的 X (其中代表误差信息的后几维主成分已被略掉). 由此得到的新样本集具有与原样本集相同的分布、相同的主特征值 (前 a 维).

2 BMPCA - ANN 建模

在 BMPCA 基础上提出了 BMPCA - ANN 的建模方法, 通过 PCA 映射和反向映射, 信息在 $\hat{X}$ 中重新分布, 由此可用于一般的 ANN 建模. 建模过程如下

$$X \xrightarrow{\text{PCA}} \sum_{i=1}^r t_i p_i^T \xrightarrow{a \leq r} \sum_{i=1}^a t_i p_i^T \xrightarrow{\text{BM}} \hat{X} \xrightarrow{\text{ANN}} Y$$

以某厂滴流床反应器建立压降模型为例, 将 BMPCA - ANN 和其他几种常用建模方法进行了比较. 通过对现有文献所发表的关联式的分析, 整理出与床层压降 Δp 相关、且可现场实测的变量, 按其相互关系分为 3 类: (1) 与反应器操作状况相关如反应器入口温度 T_{in} 、操作压力 p_{in} 、原料油流量 F_O 、混氢流量 F_H ; (2) 与原料性质相关如原料油粘度 μ 、密度 ρ 、恩氏蒸馏馏程 (American Society for Testing Materials, ASTM) 参数等; (3) 与催化剂装填情况有关如装填层数 I 、第 i 层催化剂的装填重量 G_i 、装填高度 L_i 、颗粒性质 (当量直径 d_i 、孔隙率 ϵ_i 、形状系数 φ_i 等). 其中与压降关系较大而工厂又不能提供准确数据的粘度, 利用人工神经网络的特性, 由 T_{in} 、 p_{in} 、 ρ 及 ASTM 参数进行隐式表征. 在实际生产过程中, 对压降影响还有一个重要因素是催化剂的结垢情况, 经对实际数据分析表明, 它对操作前期的压降影响不大, 而在操作后期成为主要影响因素. 本文旨在探讨用 BMPCA - ANN 方法对操作前期进行建模研究, 因此可不考虑催化剂的结垢情况, 对操作后期的建模将另文讨论. 由前分析可得到关联式

$$\Delta p = f(T_{in}, p_{in}, F_O, F_H, \rho, ASTM, G_i, L_i, d_i, \epsilon_i, \varphi_i) \quad i = 1, \dots, I \quad (13)$$

由实测数据整理得 213 组样本, 其中 191 组用于训练, 22 组用于检验. 对装填层数 $I = 2$, 原料油 ASTM 以 0、10% 及 538℃ 馏出量表示可得

$$\Delta p = f(X), X \in R^{16} \quad (14)$$

表 1 中示出了 BMPCA - ANN 建模的结果分析.

Table 1 Modeling results of BMPCA - ANN

Principle component number a	Error of training samples (rms)	Error of testing samples (rms)	number of relative error less than 20%
15	0.0177	0.0509	17
14	0.0137	0.0621	16
13	0.0153	0.0562	17
12	0.0167	0.0524	19
11	0.0232	0.0509	17
10	0.0247	0.0555	16
9	0.0286	0.0366	19

表中
$$rms = \sqrt{\frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{n-1}}$$

由以上数据可知, 在 9 维以后, BMPCA - ANN 方法建立的模型即已基本接近, 增加主成分数并不能有效改善模型的精度。

3 与其他方法的比较

3.1 PCR

PCR 方法是在 PCA 基础之上发展而来的线性回归方法之一。它利用 X 的投影 T 来代替 X , 从而得到回归公式

$$Y = TB + E \quad (15)$$

其中 B 为回归系数, E 为残差矩阵, 由于 T 具有正交性, 故 B 可容易求得

$$\hat{B} = (T^T T)^{-1} T^T Y \quad (16)$$

忽略掉小特征值所对应的坐标, 其中 T 的维数 a 取法同上。由于该方法是基于线性回归, 因此对本例非线性强的体系来说, 其回归效果较差。比较见表 2 和图 1。

3.2 PCA - ANN

非线性 PCA 是利用 PCA 分解出来的 T 矩阵作为神经网络输入, 直接映射到网络的输出, 该方法虽然也有一定的优点, 但对于主成分中含有误差时无法准确建模, 而本例中的压降问题即可能在主成分中含有较大误差, 通过实际比较可知, 该方法建立的模型精度和单独使用 ANN 方法建立的模型相同, 并无明显优势, 比较见图 1。

Table 2 Comparisons of different method

Principle component number a	PCR (rms)	PCA - ANN (rms)	BMPCA - ANN (rms)
12	0.0626	0.0666	0.0524
11	0.0614	0.0523	0.0509
10	0.0749	0.0816	0.0555
9	0.0707	0.0579	0.0366

对于采用 ANN 方法建模过程常要考虑的过拟合问题, BMPCA - ANN 也可有效去除。在单纯采用 ANN 方法建模时, 对检验样本中 T_{in} 、 p_{in} 、 F_O 、 F_H 、 ρ 5 个参数进行了趋势预测, 共

有33个趋势是相反的,分析可知,这是样本误差与ANN过拟合的结果;而采用 $\alpha = 9$ 的BMPCA-ANN时降至27个,采用 $\alpha = 12$ 时降至17个,可见BMPCA-ANN也可降低过拟合情况发生的可能性。

4 结论

本文通过探讨多元统计分析中常用的主成分分析的方法,提出了BMPCA技术,可以用于建模或分类前的数据处理。PCA方法具有寻找主成分,剔除次要因素影响的优点,BMPCA技术继承了这种优势,并改善了PCA方法中由于主成分带误差即会影响模型精度的缺点。与其他方法进行比较之后的结果表明,BMPCA-ANN网络具有较高的精度,较好的预测性能,同时降低了过拟合现象的出现,因此可用作各种工业样本建模时的参考方法之一。

参 考 文 献

- 1 Qin S J, McAvoy T J, *Computers Chem Enging*, 1992, 16(4): 379 ~ 391
- 2 李耀. 重油加氢仿真培训系统数学模型的机理及PLS建模方法的研究; [学位论文]. 北京: 清华大学, 1996
- 3 边肇祺. 模式识别. 北京: 清华大学出版社, 1988
- 4 Paul Geladi, Bruce R Kowalski. A Tutorial, *Analytica Chimica Acta*, 1986, 185: 1 ~ 17

APPLICATION OF BACK MAPPING PRINCIPAL COMPONENT ANALYSIS APPROACH IN MODELING OF CHEMICAL PROCESSES

Qin Yong, Shen Jingzhu, Hu Shanying
and Zhou Jianhua
(Department of Chemical Engineering,
Tsinghua University, Beijing 100084)

Zhou Jin
(Shengli Refinery of Qilu Petrochemical Co.,
Shandong Linzi 255434)

Abstract In this paper, a Back Mapping Principal Component Analysis method based on the analysis of the PCA is proposed. This method is compared with other modeling methods usually used in chemical processes in an example of pressure drop model for a hydrocracking reactor.

Keywords PCA, back mapping, modeling, ANN, trickle bed

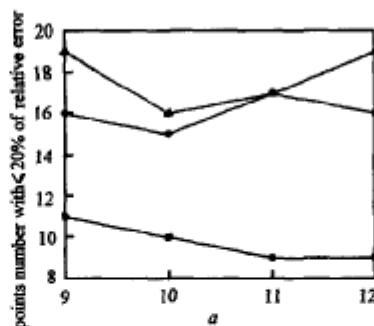


Fig.1 Comparisons of points with $\leq 20\%$ error

■ PCR; ● PCA-ANN; ▲ BMPCA-ANN